Solution Overview

NVIDIA

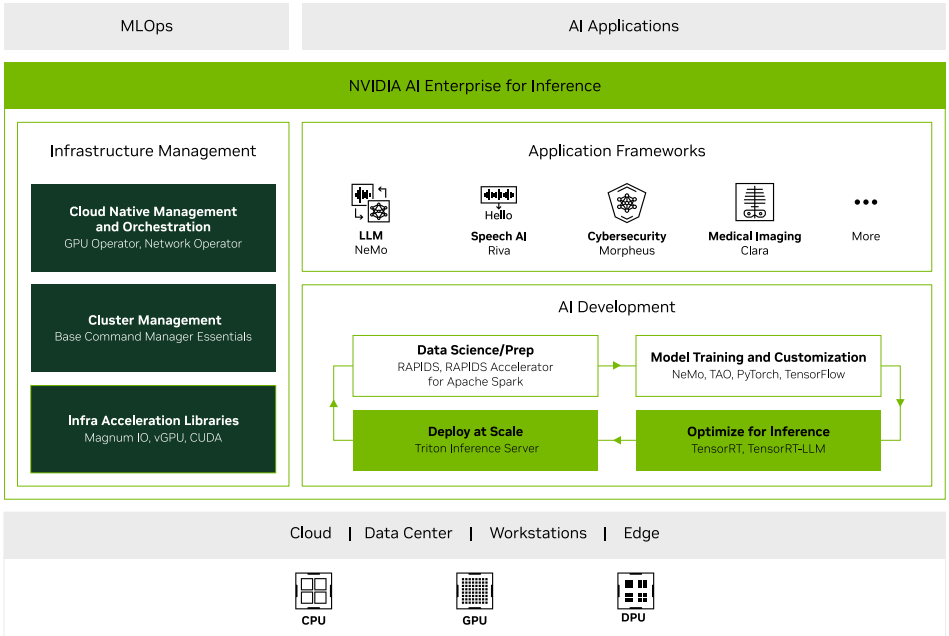
Faster, More Accurate AI With the NVIDIA Inference Platform

Drive breakthrough performance in your AI-enabled applications and services.

The Challenges of AI Inference Deployments

Deploying and running models in production applications is complex. Organizations must account for constantly evolving model frameworks with different teams using different frameworks. The computing processors, accelerators, and software platforms are also evolving at an unprecedented pace. The number and type of models are expanding as AI is making its way into many areas of business.

For the successful use and growth of AI, organizations need a full-stack approach to AI inference that supports the end-to-end AI lifecycle, with tools that help all teams reach their goals.



The NVIDIA AI inference platform workflow, from trained model to end-user application.

Deploy Next-Generation AI Inference With the NVIDIA AI Platform

NVIDIA AI offers a complete end-to-end stack and suite of products and services that deliver the performance, efficiency, and responsiveness critical to powering the next generation of AI inference—in the cloud, in the data center, at the network edge, and in embedded devices.

Key Challenges for AI in Production

- **Latency:** Real-world applications often require low latency and high throughput.
- **Integration with production data:** Different feature transformations in model training and live inference can deliver inaccuracy in production.
- **Different platforms:** Clouds (AWS, GCP, Azure, etc.), on-premises servers, mobile devices, and IoT devices are all possible locations to deploy inference.
- **Different models and machine learning frameworks:** As the models and frameworks differ, best practices for deploying and integrating also differ.
- **Scaling:** Multiple models can be involved in inference, resulting in complex prediction flows.
- **Diverse CPUs and GPUs:** Models can be executed on a CPU or GPU, and there are different types of GPUs and CPUs.

Often, organizations end up having multiple, disparate inference serving solutions—per model, per framework, or per application.

Accelerated Production AI With NVIDIA AI Enterprise

NVIDIA AI Enterprise consists of NVIDIA NIM, Triton Inference Server, TensorRT, and other tools to simplify building, sharing, and deploying AI applications. With enterprise-grade support, stability, manageability, and security, enterprises can accelerate time to value while eliminating unplanned downtime.

Key features and benefits:

- Over 100 frameworks, pretrained models, AI workflows, and development tools accelerate and simplify an enterprise's journey to AI.
- Security and API stability are ensured for interdependent software packages within the inference software stack.
- Software optimizations increase performance and lower TCO.
- Infrastructure and inference management are simplified with NVIDIA Base Command™ Manager Essentials.
- Industry ecosystem certifications are available across the cloud, data center, workstations, and edge.

The software platform comes with enterprise support with service-level agreements (SLAs) and access to AI experts.

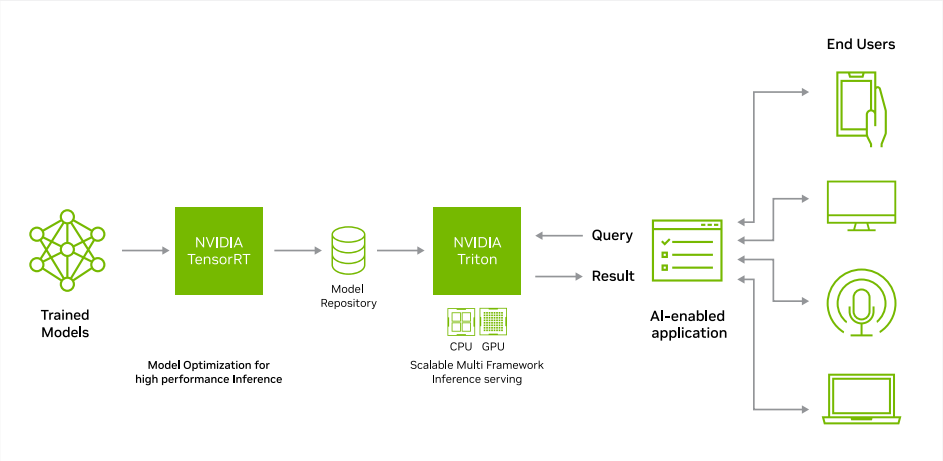


Diagram displaying how NVIDIA's AI inference software works.

NVIDIA Triton

NVIDIA Triton Inference Server is open-source inference serving software that standardizes AI model deployment and execution from all major AI frameworks on any GPU- or CPU-based infrastructure.

Key features:

- High performance and utilization are achieved on both GPU and CPU systems through request batching and concurrent model execution.
- Stringent application latency SLAs for real-time and offline batched inference and large language models (LLMs) come with support for multi-GPU, multi-node execution and model ensembles.
- PyTriton provides a simple interface that lets Python developers use Triton Inference Server to serve anything, be it a model, a simple processing function, or an entire inference pipeline.

The End-to-End NVIDIA AI Inference Platform

- **NVIDIA AI Enterprise:** An end-to-end AI software platform consisting of NVIDIA Triton™ and TensorRT™ to simplify building, sharing, and deploying AI applications. With enterprise-grade support, stability, manageability, and security, organizations can accelerate time to value while eliminating unplanned downtime.
- **NVIDIA Triton Inference Server:** An open-source inference serving software that's key for fast and scalable inference in every application.
- **NVIDIA TensorRT:** An SDK for high-performance deep learning inference, key for delivering low latency and high throughput to inference applications.
- **NVIDIA TensorRT-LLM:** An open-source library enabling developers to accelerate and optimize inference performance of the latest LLMs on NVIDIA GPUs.
- **NVIDIA NIM™ microservices:** A set of easy-to-use microservices designed for secure, reliable deployment of high-performance AI model inferencing across clouds, data centers, and workstations.
- **Wide ecosystem integration:** The NVIDIA AI inference platform is tightly integrated into and works with all major platforms, including Amazon Sagemaker, Azure Machine Learning, Google Vertex AI, TensorFlow, Pytorch, and more.
- **Accelerated computing infrastructure:** Hardware systems designed for optimizing AI workloads.

- Triton standardizes AI model deployment for all applications across cloud and edge, and it's in production at world-leading companies—Amazon, Microsoft, American Express, and thousands more.
- Triton's model analyzer can shrink model deployment time from weeks to days. It helps select the optimal deployment configuration to meet the application's latency, throughput, and memory requirements.

NVIDIA TensorRT

NVIDIA TensorRT, an SDK for high-performance deep learning inference, includes a deep learning inference optimizer and a runtime that deliver low latency and high throughput for inference applications. TensorRT can be deployed, run, and scaled with Triton.

Key features:

- Built on the NVIDIA CUDA® parallel programming model, TensorRT optimizes techniques such as quantization, layer and tensor fusion, and kernel tuning on NVIDIA GPUs.
- TensorRT provides INT8 using quantization-aware training and post-training quantization and FP16 optimizations for deployment of deep learning inference applications, such as video streaming, recommendations, anomaly detection, and natural language processing.
- Integrations with all major frameworks, including Pytorch and TensorFlow, allow users to achieve 6x faster inference with a single line of code.

NVIDIA TensorRT-LLM is an open-source library that accelerates and optimizes inference performance of the latest LLMs on NVIDIA GPUs. It enables developers to experiment with new LLMs, offering speed-of-light performance with quick customization capabilities without deep knowledge of C++ or CUDA optimization.

TensorRT-LLM wraps TensorRT's deep learning compiler, optimized kernels from FasterTransformer, pre- and post-processing, and multi-GPU/multi-node communication in a simple open-source Python API for defining, optimizing, and executing LLMs for inference in production.

NVIDIA NIM

Part of NVIDIA AI Enterprise, NVIDIA NIM is a set of easy-to-use microservices designed to accelerate deployment of generative AI. These prebuilt microservices support a broad spectrum of AI models—from open-source community models to NVIDIA AI Foundation and custom models. NIM microservices can be deployed with a single command and quickly integrated into applications with just a few lines of code.

Key features:

- Built on robust foundations, including inference engines like Triton, Inference Server, TensorRT, TensorRT-LLM, and PyTorch, NIM facilitates seamless AI inferencing at scale.
- NIM ensures that AI applications can be deployed anywhere with confidence. Whether on premises or in the cloud, NIM is the fastest way to achieve accelerated generative AI inference at scale.
- The latest community-built AI models—optimized and accelerated by NVIDIA—are available at ai.nvidia.com. With API access to these models, developers can experiment, prototype, and ultimately deploy anywhere with NVIDIA NIM.

Key Benefits of NVIDIA AI Inference

- Standardized deployment: Standardize model deployment across applications, AI frameworks, model architectures, and platforms.
- Easy integration: **Integrate easily** with tools and platforms on public clouds, in on-premises data centers, and at the edge.
- **Lower cost:** Achieve high throughput and utilization from AI infrastructure, thereby lowering costs.
- **Seamless scalability:** Scale inference jobs seamlessly across one or more GPUs.
- **High performance:** Experience incredible performance with the NVIDIA inference platform, which has set records across multiple categories in MLPerf, the leading industry benchmark for AI.

NVIDIA Accelerated Computing Infrastructure

NNVIDIA offers a complete infrastructure portfolio for AI workloads, including NVIDIA GPUs, the NVIDIA Blackwell architecture, NVIDIA Grace™ CPUs, NVIDIA GB200 Grace Blackwell Superchips, NVIDIA DGX™ B200, NVIDIA GH200 Grace Hopper™ Superchips, and accelerated networking with NVIDIA BlueField® DPUs.

Key features:

- A complete GPU portfolio with the NVIDIA Blackwell and NVIDIA Hopper™ architectures, from entry level to mainstream to the highest performance. Each GPU has the versatility to accelerate AI applications, whether at the edge, in the cloud, or on premises.
- The NVIDIA AI software stack is designed to work with all major x86 CPU systems, including AMD and Intel CPUs, and is optimized for the Grace CPU, GB200 Superchips, and GH200 Superchips to accelerate AI and high-performance computing applications.
- The BlueField DPU platform offloads, accelerates, and isolates a broad range of advanced infrastructure services, providing AI data centers with high performance, robust security, and sustainability.
- NVIDIA Quantum InfiniBand and NVIDIA Spectrum™ Ethernet networking platforms provide AI practitioners with advanced acceleration engines and the fastest speeds at 400 gigabits per second, enabling superior performance for inference at scale.
- Designed for industry-leading performance and multi-node scale with NVIDIA DGX BasePOD™ and DGX SuperPOD™, NVIDIA DGX systems are the gold standard in AI infrastructure, delivering the fastest time to solution on the most complex AI workloads.
- NVIDIA-Certified Systems™ bring NVIDIA GPUs and NVIDIA high-speed, secure network adapters to systems from leading NVIDIA partners in configurations validated for optimum performance, manageability, and scale.

Ready to Get Started?

To learn more about the NVIDIA AI inference platform, visit: nvidia.com/inference

To contact us about purchasing our AI inference software, visit: nvidia.com/ai-enterprise-sales